Impact Evaluation Methods in Public Economics: A Brief Introduction to Randomized Evaluations and Comparison with Other Methods

HARVARD | BUSINESS | SCHOOL

Impact Evaluation Methods in Public Economics

Dina Pomeranz

Working Paper 16-049

Impact Evaluation Methods in Public Economics

Dina Pomeranz
Harvard Business School

Working Paper 16-049

Impact Evaluation Methods in Public Economics

A Brief Introduction to Randomized Evaluations and Comparison with Other Methods*

Dina Pomeranz

September 2015

Recent years have seen a large expansion in the use of rigorous impact evaluation techniques. Increasingly, public administrations are collaborating with academic economists and other quantitative social scientists to apply such rigorous methods to the study of public finance. These developments allow for more reliable measurements of the effects of different policy options on the behavioral responses of citizens, firm owners, or public officials. They can help decision makers in tax administrations, public procurement offices, and other public agencies design programs informed by well-founded evidence. This paper provides an introductory overview of the most frequently used impact evaluation methods. It is aimed at facilitating communication and collaboration between practitioners and academics by introducing key vocabulary and concepts used in rigorous impact evaluation methods, starting with randomized controlled trials and comparing them with other methods ranging from simple pre-post analysis to difference-in-differences, matching estimations, and regression discontinuity designs.

Keywords: public economics, taxation, impact evaluation, randomized controlled trials

JEL codes: H2, H3, C93, C21

* I thank Michael Eddy and Stephanie Majerowicz for excellent research assistance and the officials of the Chilean and Ecuadorian tax authorities, the Chilean procurement authority, and the Chilean National comptroller agency for helpful feedback and suggestions.

1. Introduction

Daily decisions made in public finance can affect the economy of an entire country. However, assessing the effectiveness of different policy options is challenging. Public officials are constantly confronted with a myriad of important questions related to the impacts of public policy on the behavior of citizens, firms, or public officials. What policies are most effective against tax evasion? How strongly will firm owners react to tax incentives? How can monitoring be optimized to improve the behavior and compliance of public procurement officials? What type of communication can motivate school officials to disperse educational infrastructure grants promptly? How is the design of optimal policies affected by the behavioral biases that have been identified by the growing behavioral public economics literature (Chetty 2015)?

Recent improvements of impact evaluation techniques allow for increasingly reliable answers to these types of questions. A growing number of collaborations between public administrations and academics have facilitated the application of randomized evaluations and other quasi-experimental methods to questions of public finance and behavioral economics. In public economics, impact evaluations can often take advantage of already available administrative data, which considerably reduces their cost.

There are various impact evaluation methods, each with different degrees of validity. The quality of the evaluation is of utmost importance for obtaining informative, unbiased results. This paper provides an overview of the most frequently used methods, in a language that is accessible both to academics and practitioners in public economics. It offers a brief summary of each method, its advantages and drawbacks, and the conditions under which the method produces valid results. In addition, it provides an introduction to key elements of the specialized terminology of impact evaluations in order to facilitate the communication between policymakers and academics looking to collaborate on these topics.

It is, therefore, useful to define some basic concepts before presenting the specific methods below. The objective of every impact evaluation is to demonstrate a ***causal effect.*** The goal is to measure the impact of a program or policy on some outcome of interest. For example, what is the effect of a notification letter on tax payments? In the context of impact evaluations, the policy or program whose impact we want

to analyze is often referred to as the ***treatment.*** The impact is then the result that can be attributed directly to the treatment – such as a change in tax filings as a result of the notification letter.

The fundamental challenge of impact evaluation is that at any given moment, it is only possible to observe what happened given the policies in place, not what would have occurred without those policies. It is possible to observe tax filings of taxpayers that received a notification, but it is not possible to observe what those same taxpayers would have done in the absence of such a notification. This imaginary situation of what would have happened in the absence of the treatment is called the ***counterfactual*****.** Understanding the counterfactual is key to understanding the impact of a program. Figure 1 provides a graphical representation of this unobserved counterfactual.

[Figure 1]

Figure 1 represents the fundamental challenge of impact evaluations, which seek to measure the difference between the outcome that in fact occurred (shown in light/yellow dots) and the counterfactual that is never observed (shown with dark dots). In this example, we can see that the primary outcome increased more steeply after the intervention (light dots) than would have been the case without the intervention (dark dots). The impact is measured as the difference between the outcome with the treatment and the outcome that would have happened without the treatment (the counterfactual).

If an accurate representation of the counterfactual existed, then impact evaluation would be easy. The impact of a program or policy would be the difference between the result observed with the program and the result that would have prevailed without the program. Given that the counterfactual can never be observed in reality, each evaluation tries – in an explicit or implicit manner – to construct an estimate of the counterfactual to compare it to what occurred. The quality of that representation drives the quality of the impact evaluation.

Normally, the counterfactual estimate is represented by a group called the ***control group*** or ***comparison group*****.** The control group consists of people or firms that did not participate in the program, while

the ***treatment group*** is the group that participated in the program. To measure the impact of the intervention, the outcomes of the treatment group are compared with the outcomes for the control group. An evaluation will produce reliable results if the control group is identical to the treatment group in all its characteristics – observable or not – except one: their exposure to the treatment. In this case, any difference after the intervention can be attributed to the program. In the absence of treatment, both groups would be the same, so the control group constitutes a valid representation of the counterfactual.

All methods used to construct the comparison group rely on assumptions under which the two groups would be comparable. When the assumptions are realistic, the control group is a good representation of the counterfactual. When these assumptions are not realistic, the resulting impact evaluation will be ***biased.*** That means it may over- or under-estimate the true effect. A biased evaluation may result in poorly-informed policy decisions and generate losses in terms of effort, time, and public resources. It is, therefore, important to use high-quality methods in order to obtain a reliable impact estimation, and to provide solid evidence for decision-making.

Bias can stem from a variety of reasons that make the treatment and comparison groups different. ***Selection bias*** is produced when those selected into the treatment group are different from those in the comparison group in a way that affects outcomes. This happens also when people who choose to participate in a treatment are different from those who do not (self-selection). Bias can also come about when an external factor affects those in the treatment differently from those in the comparison group. This is sometimes referred to as ***omitted variable bias.*** It biases the conclusion that is reached by comparing the treated group to a comparison group that no longer represents a valid counterfactual**.**

The focus on making the estimation accurate and unbiased is known as ***internal validity.*** Internal validity indicates the extent to which a causal conclusion based on a study is warranted, i.e., the extent to which a study avoids the risk of bias. Well-executed randomized evaluations have very high internal validity. Other methods described below have larger risks of bias, and consequently, lower internal validity. These will be discussed in more detail below.

In contrast, ***external validit***y refers to the extent to which the causal findings of a study can be generalized or extrapolated to other situations and settings. For instance, in the area of public economics, an external validity question could ask to what extent the findings of an evaluation in one region are informative for a potential nation-wide rollout of a policy, or even for other countries or continents. External validity can to some degree be assessed based on specific knowledge of the setting in question, or one can explicitly test for it through replication of the same analysis in different settings. See Banerjee and Duflo (2009) and Duflo, Glennerster and Kremer (2008) for a more in-depth discussion.

The remainder of the paper discusses characteristics, strengths, and limitations of different evaluation methods (for a more in-depth treatment of any of these methods, see, for example, Angrist and Pischke 2009, 2015; Imbens and Wooldridge 2009; and Gertler et al. 2011). Section 2 starts with randomized evaluations as the benchmark to which the other methods can be compared. Sections 3-4 discuss simple difference and simple pre-post analysis. These methods require the strongest assumptions and are most likely to yield biased results. Sections 5-6 present difference-in-differences analysis, matching procedures, and propensity scores. Depending on the setting, these methods can yield reliable impact estimations, but they have to be applied selectively and with great care to ensure their underlying assumptions are met. Section 7 provides an introduction to the regression discontinuity design. This method can, under certain circumstances, deliver causal estimates that are just as valid as those from randomized evaluations, with the caveat that they estimate the effect only for a specific subsection of the population. Section 8 concludes.

2. Randomized Evaluation

The goal of ***randomized evaluations*** – also called ***experimental evaluations, randomized controlled trials (RCTs),*** or ***randomized field experiments*** – is to create an ideal comparison group by design from the beginning of the intervention. Study participants, which can be individuals, firms, or entire public entities or localities, are randomly assigned to either receive the treatment or be in the comparison group. This random assignment ensures that (on average) there is no difference between the individuals in

the treatment and control group, except for the fact that one group has been randomly chosen to participate in the program and the other has not. We can therefore rule out that the impact measured is due to a systematic difference between the treatment and control group that would have existed even without the application of the treatment (Duflo, Glennerster and Kremer 2008). Randomized evaluations are thus often seen as the ideal way to conduct an impact evaluation. It is for this reason that in the evaluation of new medicines and in natural science laboratory research, this method is used almost exclusively. (Note: It is important to distinguish between a randomized evaluation and a random sample. Random samples are used to obtain representative descriptive information about a population, not to measure impact. The distinctive characteristic of a randomized evaluation is that the treatment is assigned randomly.)

Another benefit of randomized evaluations is that they allow researchers to identify the effect of a particular component of a larger program. To do so, one can vary one particular factor in a number of treatment groups and compare them to the control group. This way, the casual impact of a particular component of a program or policy can be identified in a way that is difficult otherwise (Banerjee and Duflo 2009). For instance, studies about what policies can improve access to education and school learning sought to measure the specific effects of textbooks (Glewwe, Kremer and Moulin 2009), class-size (Angrist and Lavy 1999), and student health (Miguel and Kremer 2004). Randomized evaluations that manipulate one factor at a time, while holding the other elements of the classroom environment constant, can measure the individual impact of each factor. This isolation of specific factors can make it possible to test particular mechanisms through which a policy has an effect (Ludwig, Kling and Mullainathan 2011).

Importantly, randomized assignment requires that the evaluation be designed before the program has begun. For this reason, this method is also called ***prospective evaluation*****.** In a random process, individuals (or other entities like schools, firms, or villages) are assigned to the treatment group and those not selected are part of the control group. This generates two groups that are similar both in terms of observable characteristics (such as education levels) and unobservable ones (such as motivation). Therefore, any difference that arises later between the treatment and control groups can be attributed to the program

and not to other factors. For this reason, if designed and applied adequately, a randomized evaluation is the most valid method for measuring the impact of a program and requires the fewest additional assumptions.

2.1 *Randomization in practice*

This section will lay out a brief overview of the different steps involved in setting up and implementing a randomized field study (for a more detailed description of the steps involved in randomized controlled trials under many different scenarios, see Glennerster and Takavarasha 2013, and Gerber and Green 2012). The first step is to choose a program, population, and main outcome variables of interest. Ideally, this will be a program that is of interest to the policymaker to the degree that learning about its effectiveness, or aspects of its effectiveness, will feed into the decision-making process of the public entity.

Second, prior to starting the evaluation, it is useful to calculate statistical estimates to determine the size of the treatment and control groups required for reliably measuring the impact on outcome variables of interest. This analysis is called ***power calculation*** since it estimates how many observations are needed to have enough statistical power to detect a meaningful effect.

How do we determine the number of participants required in a randomized study? The greater the number of individuals included in a study, the more likely it is that both groups will be similar (due to the statistical "law of large numbers"). This is one of the reasons why sample size is important. A larger sample is better since it reduces the likelihood of having unbalanced groups. Moreover, a larger sample improves the precision of the impact estimates, i.e., it increases the likelihood of detecting the true impact of a program. Nevertheless, a bigger study can be more costly and is not always feasible. Therefore, power calculations can help determine the sample size necessary for measuring the impact on the main outcome of interest.

Statistical power calculations incorporate the different factors that affect the number of required participants. Among the factors to be considered are the variance of the outcome variable of interest and the minimum effect expected to be detected. The smaller the size of the effect one wishes to detect, the larger the number of observations needed. In addition, the higher the variance in the outcome of interest,

the larger the number of observations needed to distinguish the true effect size from simple noise in the data. Finally, the randomization design can affect the necessary group size. If the randomization is performed at the group level (clustered randomization), more observations will be necessary than if the randomization is done at the individual level (see more details on clustered randomization below.)

The third step in a randomized evaluation is the random assignment of treatment. The randomization process can be as simple as tossing a coin or a conducting a lottery. To make the process transparent and replicable, the random assignment is often implemented using a statistical software such as Stata. It is important that the randomization process be truly random and not just "seemingly" arbitrary. For example, assigning the treatment to people whose surnames start with the letters "A-L" and leaving those starting with "M-Z" as control may seem random, but it is not. Such assignment requires the assumption that the individuals whose surnames start with the letters "A-L" are the same as those that start with "M-Z". However, it is possible that the families whose surnames start with the letters "A-L" are different from the families with a last name starting with the letters "M-Z". For example, the ethnic composition may vary. To avoid this situation, an automated method such as using a computer program to generate random numbers that determine treatment assignment is recommended.

A computer also simplifies more complex randomization processes, like ***stratified randomization*****.** Stratified randomization is recommended when the number of potential participants is relatively small, to ensure that both groups are balanced with respect to the most important variables. In stratifying, the sample is divided into subgroups of similar characteristics, with participants within each subgroup randomized to treatment and control, such that the proportion in treatment and control is the same for each subgroup. For example, if the population is divided by gender, if thirty percent of men and thirty percent of women are assigned the treatment, this assignment will be perfectly balanced in terms of gender. The treatment group will have the exact same gender composition as the control group.

As mentioned above, another often-used randomization design is ***clustered randomization*****.** In this procedure, the randomization is not conducted at the level of an individual, but at the level of groups of

individuals: entire groups (or "clusters") of people are assigned to either treatment or control. This is particularly useful for situations in which it can be expected that the treatment will have spillover effects on others in the same group. For example, when testing the effect of a new textbook, random assignment at the student level may not be possible, as the teacher will be teaching from the same book to the entire class. The assignment should then be done at the class level. Another example could be a tax authority that wants to test a new communication strategy towards small firms. They might worry that tax accountants, which work for several firms, could share information across the firms they work for. To remedy this, the randomized assignment could be done at the accountant level, such that firms that share the same accountant would either all be in the treatment group or all in the control group.

It is not necessary for both groups to be the same size. However, it is important to verify that the groups are balanced with respect to the main outcome variables of interest. That is, the average characteristics (e.g., average firm revenue, industry composition, or percent of women) are not significantly different between the treatment and the control group. In the academic literature, experimental studies therefore usually include a balance table that shows that the main characteristics are similar across the two groups.

The fourth step in a randomized evaluation should – whenever possible – be a pilot phase of the planned intervention. A small-scale pilot implementation of the program to be evaluated can provide enormous benefits for the preparation of the large-scale intervention. In practice, the lessons learned from the pilot often make the difference between a successful, informative randomized study and an unsuccessful one. Pilots allow researchers and policymakers to learn about unforeseen challenges at a small scale, when they can still be remedied, and avoid unexpected problems later. This applies both to the implementation of the program itself, as well as the data collection process, the internal communication in the public agency about the intervention, etc. This logic of piloting and testing the intervention before conducting the large-scale program evaluation is also consistent with practices used in Silicon Valley style technology start-up environments, where it is often known as the "Lean Startup" approach (Ries 2011).

Finally, the implementation of the program or policy to be evaluated is carried out. The most important aspect of this process is to make sure that there is no difference between the treatment and control

group except the application of the program. Sometimes, well-meaning officials misunderstand the idea of the control group and think that all other interventions towards the control groups should also be halted until the study ends. However, this would amount to treating the control group differently from the treatment group. For instance, imagine a tax authority wants to test a new communication strategy by sending specific letter messages to a randomly selected group of taxpayers and comparing their behavior to a control group. If officials now decided to halt all auditing activities in the control group but continue to apply such audits to the treatment group (or vice versa), the validity of the study would be lost. In this case, the two groups would not only differ in terms of receiving the treatment, but also in terms of their risk of being audited. When looking at the final difference between the two groups, it would be impossible to establish whether the difference stems from the treatment or from the effects of the audits.

During the implementation, it is also important to make sure that the random assignment of individuals to each group is respected and that participants are not moved from one group to another. In the event that the randomization is not respected in the implementation process, it is still possible to conduct a valid impact evaluation, as long as researchers have precise information about who ended up receiving the treatment and who did not. In this case, it is possible to use the "Intent-to-Treat" methodology and use instrumental variables to measure the "Treatment-on-the-Treated" effect. This approach could, for example, be used if some letters sent to taxpayers were not received due to incorrect addresses (as done in Pomeranz 2015). It is very important that even if this happens, the original random assignment is used when conducting the impact evaluation; those that were *assigned* to the treatment have to be compared to those *assigned* to be in the control group. It is never valid to compare those who were in fact treated with those that were meant to be treated, but ultimately did not participate in the program, because these two groups will not be comparable. In our example, taxpayers for whom the tax authority has invalid addresses are likely to be systematically different in many aspects from those with valid addresses.

2.2 *Experiences of randomized evaluations in public economics*

Recent years have seen a strong increase in the use of randomized field experiments to study many different areas of public policy. One such area is tax administration. A pioneering collaboration of this nature was undertaken by Coleman (1996), Blumenthal, Christian and Slemrod (2001), and Slemrod, Blumenthal and Christian (2001) undertook a pioneering collaboration of this nature with the tax authority of Minnesota in the mid-1990s. Many academics have followed their example, and a growing number of tax authorities are collaborating with academics. Randomized experiments have since been conducted by tax authorities in Argentina, Australia, Austria, Chile, Denmark, Ecuador, Finland, Germany, Israel, Mexico, Peru, Switzerland, USA, Venezuela (Hallsworth 2014), and plans for such projects are under way in Kenya, Liberia, Rwanda, Uganda, and other countries around the world.

One frequently used type of intervention consists of sending letter messages to taxpayers in order to test different hypotheses about taxpayer behavior. The most frequently used outcome measures relate to the amount of taxes paid, since tax administrations already have access to this data; it is the first order of concern for tax administrations. A growing number of recent studies have measured the impact of randomized letters or text messages on the behavior of individual taxpayers (Coleman 1996; Blumenthal, Christian and Slemrod 2001; Slemrod, Blumenthal and Chrisitan 2001; Torgler 2004, 2013; Wenzel 2005, 2006; OECD 2010; Kleven et al. 2011; Fellner, Sausgruber and Traxler 2013; Haynes et al. 2013; Dwenger et al. 2014; Hallsworth et al. 2014; Bhargava and Manoli 2015), property owners (Wenzel and Taylor 2004; Del Carpio 2013; Castro and Scartascini 2015), or firms (Hasseldine et al. 2007; Iyer, Recker and Sanders 2010; Ariel 2012; Harju, Kosonen and Ropponen 2013; Ortega and Sanguinetti 2013; Bhargava and Manoli 2015; Pomeranz 2015). Some letters have tested behavioral responses to either audit threats or motivational messages. Others have evaluated the importance of the wording, such as the simplicity and clarity of the message (Bhargava and Manoli 2015). Other studies include additional measures such as face-to-face visits (Gangl et al. 2014). For an excellent overview on the use of randomized field experiments to increase tax compliance, see Hallsworth (2014).

In collaboration with the tax authority in Chile, we employed this type of randomized letter message experiment for a particularly policy-relevant aspect of tax administration: risk indicators that predict what

types of taxpayers are more likely to react to an increase in the audit probability (Pomeranz, Marshall and Castellon 2014). Many tax authorities use such risk indicators to select which taxpayers will be audited. However, inputs into such risk indicators often suffer from a self-fulfilling circle problem. Information about high evasion is typically found through audits. This information is therefore more available from types of taxpayers that were already audited more frequently in the past. The risk indicators therefore end up having a self-referential problem, in which types of taxpayers that were audited more in the past are more prone to be found as high risk in the future. We developed a method that gets around this problem, by using the response to randomized deterrence letter messages as inputs into the risk indicator. Tax authorities can apply this method to target audit activities towards categories of taxpayers that can be expected to have a particularly strong response.

In addition to analyzing the impacts of different communication and auditing strategies, randomized studies can also be used to study behavioral responses of taxpayers to the tax structure itself. In collaboration with the Chilean tax authority, we evaluated the role of third-party information for value added tax (VAT) compliance (Pomeranz 2015). The results show that the VAT can indeed have important "self-enforcing" properties. However, these properties are only activated if the audit probability is high enough that taxpayers take the risk of detection seriously. In this case, the third-party information can lead to important spillover effects that multiply the effectiveness of tax enforcement measures.

Taxation is by no means the only area of public economics in which randomized experiments play a growing role. Public procurement is another area of growth for these types of studies. Projects are currently under way in procurement agencies in Brazil, Chile, and Colombia among others. One of the few randomized studies in this area that has already been completed is Litschig and Zamboni (2013). They analyze whether a randomized increase in the audit risk deters corruption and waste in local public procurement in Brazil. The results show that a twenty percentage point increase in the audit risk reduces the incidence of corruption and mismanagement of local procurement by seventeen percentage points.

Governments may also want to study many other aspects related to the effectiveness of government spending. For example, in the area of savings, randomized evaluations in very different settings found (by

randomly varying the savings interest rate) that subsidizing interests rates to encourage the poor is not very effective, but that follow-up and feedback messages may be more impactful (Karlan et al. 2010; Karlan and Zinman 2014; Kast, Meier and Pomeranz 2014). This suggests that the barriers to savings may be more behavioral than financial, so that inexpensive interventions such as setting defaults (Madrian and Shea 2001; Choi et al. 2002) or sending follow-up messages can be highly effective. This can have important impacts for those affected. Studies that provided randomly selected low-income individuals access to free savings accounts found that they can help the poor cope with economic shocks (Kast and Pomeranz 2014), increase monetary assets (Prina 2015), and increase investments in health and education (Dupas and Robinson 2013; Prina 2015).

There is also a large literature using randomized evaluations in the areas of public health, education, etc. Providing an overview on these areas goes beyond the scope of this paper. The website of the Abdul Latif Jameel Poverty Action Lab, http://www.povertyactionlab.org, provides a list of such evaluations conducted by its affiliates.

2.3 *Summary on randomized evaluations*

Randomized evaluations allow estimating the effect of a program or policy on the behavior of those affected by it. The fact that participants are randomly assigned to treatment makes it possible to measure the effect by simply comparing the outcomes of those assigned to the treatment group and those assigned to the control group (also called "comparison group"). The counterfactual for the treatment group is represented by the control group. Members of the treatment and comparison group are selected randomly before the start of the program, among a group of potential participants. Estimates obtained through randomized evaluations have extremely high internal validity. They require very few additional assumptions to be valid. For these reasons, randomized evaluations are often referred to as the "gold standard" in impact evaluations. The key assumption of this method is that the randomization process is executed correctly. If that is the case, the treatment and comparison groups are in expectation statistically identical along both observable and unobservable characteristics. In addition, it is important that no other treatment is applied to only one

group and not the other. One practical drawback is that the random assignment has to be done before the program is implemented, and as a result, it is not possible to carry out retrospective randomized evaluations. In addition, in certain cases, random assignment to a particular treatment may not be practically, politically, or ethically feasible.

The following sections describe other evaluation methods that try to construct an approximation of the counterfactual in circumstances where randomization is not possible. The validity of each method will depend on how similar the treatment group is to the control group.

3. Simple Difference: Comparing the Treated to the Untreated

The ***simple difference*** is one of the most frequently used methods employed to describe impacts. However, in many circumstances, its application will not provide correct, unbiased results. This section describes how simple differences work and what assumptions need to hold for them to be valid. Understanding the limits of simple differences will also further illustrate the benefits of having a valid comparison group in order to be able to obtain unbiased impact evaluations.

The simple differences methodology is straightforward: comparing the group that received the program with another that did not. The comparison group in this case corresponds to people or entities that did not participate in the program. That is, the assumption is that those who did not participate represent a valid counterfactual of what would have happened to those who received the program, had they not received the program. Unfortunately, in many cases, this assumption is not realistic. In many programs, there is a selection process that determines who receives the treatment. For example, consider an audit program in which only taxpayers identified as high risk are selected. This assignment is not random and introduces selection bias. In other cases, anyone can participate in a treatment, but people self-select if they want to participate. There is likely to be a difference between those who did or did not participate, for example, in terms of their motivation or their needs.

To illustrate this situation with a concrete example, suppose someone wants to measure the impact of a program that offers free tutoring for children who have difficulty in school. This was the case in the study by Banerjee et al. (2007), which evaluated the effect of offering separate classes to the weakest students. In these remedial classes, young women tutored students (so-called Balsakhi) in basic reading, writing, and math to help them catch up with their peers. If this study simply compared the grades of children that received help from a tutor with those that did not, the results would be misleading. It is very well possible that the children with tutors would be found to have lower grades than those without tutors. However, concluding, based on this observation, that the tutors hurt the academic achievement of the children would likely be erroneous. In this program, children who had fallen behind were selected for the remedial classes. So children who had lower grades were more likely to receive the help of a tutor. This introduces a selection bias. In this case, the selection bias leads to an underestimate of the impact. Because the treated group had lower grades to begin with, when comparing those that receive the help of a tutor to those who do not, it may appear that the tutoring had a negative effect on grades.

Despite the potential serious concerns with selection bias, simple differences are often popular because they can be conducted in a retrospective manner, even after the program has been concluded, and they do not require a lot of data (for example, no data on the situation of the participants prior to the program start). Therefore, newspapers and government documents frequently report such differences as evidence for the benefit (or lack of benefit) of certain programs. Based on the discussion above, such statements have to be treated with much caution.

3.1 *Summary on simple differences*

Analysis based on simple differences measure the impact by comparing the post-treatment situation of those that participated in a program with a comparison group that did not. The counterfactual is represented by those in the comparison group. The key assumption of this method is that those in the comparison group are identical to those that participated in the program, except for the effects of the program. A key advantage, and reason for its frequent use, is that this method does not require data on the situation prior to

the treatment. However, a big drawback is that if the treated and comparison groups are different in any way prior to the program, the method may be biased and may under- or overestimate the real impact of a policy; that is, selection bias is introduced into the estimation.

4. Pre- vs. Post-Treatment Comparison

A ***pre-post comparison*** is a particular type of simple difference evaluation. Instead of using another group as a control group, the same group of people is compared before and after participating in the program. Therefore, a pre-post evaluation measures change over time. The impact is measured as the difference between outcomes of interest before and after an intervention. The pre-post analysis is frequently used in evaluating programs. In many cases, when there is data on outcomes prior to the intervention, this type of retrospective analysis seems convenient, particularly because it does not require information on people who did not participate in the program.

In the aforementioned example of a tutoring program, a pre-post evaluation would allow taking into account the initial grades of the students. However, the important question to assess the validity of a pre-post evaluation is the following: is the situation of the participants before the start of the program a good representation of the counterfactual? In other words, is it correct to assume that without the program, during this period, there would have been no change in the results of the treated group? Figure 2 represents this issue graphically.

[Figure 2]

In the free tutoring program example, it is very unlikely that the children would not have improved their learning at all over time, even in the absence of a tutor. However, a simple pre-post evaluation would assume that all improvements over the time span of the program are due to the program. So even the learning resulting from the normal development of the children would be attributed to the tutoring program. In

other words, the estimates would have a positive bias: they would overestimate the true effect of the program.

In addition to such overall time trends, the results of a pre-post analysis can also be biased due to other factors that change the outcome over time but are not related to the program. For example, if there is an economic crisis during the implementation period of an auditing program, tax behavior may change independently of the auditing program. It is then not possible to know if the change over time is due to the crisis, the policy, or a combination of both. That is, the evaluation may be affected by omitted variable bias.

4.1 *Experiences of pre-post comparison evaluations in public economics*

While a simple pre-post comparison will often lead to biased results, there are certain settings in which a pre-post analysis can yield credible estimates, i.e., settings in which the pre-treatment situation provides a valid counter-factual for the post-treatment situation. One such example is Carrillo, Pomeranz and Singhal (2014). In this study, we evaluate a program by the Ecuadorian tax authority. The program focused on firms whose declared revenue was much lower than information about the firms' revenue that the tax authority obtained from third-party sources. Several years after the corresponding tax filings, the tax authority sent letters to firms with a particularly large discrepancy, asking them to amend their declaration. This lead to an immediate spike in the amendment rate, while firms that did not receive a letter were very unlikely to make any amendments such a long time after the original filing. In this case, a valid counterfactual for the new amount declared in the amendment is the amount declared in the original tax filing. The underlying assumption in this case is that in the absence of the letter, these firms would not have filed an amendment at this time. The study found that when firms were notified about detected revenue discrepancies, they increased reported revenues – but also reported costs, leading to only minor increases in tax collection.

4.2 *Summary on pre-post comparison*

Pre-post analysis measures the change in outcomes over time for participants of a program. It compares the situation before and after a treatment. The counterfactual is represented by the same participants, but prior to the program. The key assumption of this method is that the program is the only factor that influenced a change in outcomes over that time period. Without the program, the outcomes would have remained the same. This is, in reality, only rarely the case. Many factors that vary over time can affect an outcome, which contradicts the key assumption made above. In particular, the pre-post comparison does not control for general time trends or other things that happen over the study period that are unrelated to the program but affect the outcomes. The benefit of this method is that it does not require information on people that did not participate in the program. This is why it is often used by the media and in policy communications.

5. Difference-in-Differences Estimation

A ***difference-in-differences*** evaluation combines the two previous methods (simple difference and pre-post) to take into account both the differences between the two groups and changes over time. The effect is calculated by measuring the change over time for the treated group and the comparison group and then taking the difference between these two differences (hence the name "difference-in-differences").

[Table 1]

Table 1 shows a numerical illustration of a difference-in-differences estimation for the tutoring example. It displays the average grades of the children with and without the tutoring program, before and after the program (on a scale of zero to a hundred). As we can see, the treated group that receives a tutor has lower grades than the untreated group, both before and after the treatment. So a simple difference would have introduced a negative bias into the analysis. The numbers also illustrate that the grades of both groups improved over time. So a simple pre-post analysis would have introduced a positive bias. When we take

the difference between the two differences, we see that the grades of those who received a tutor improved by 6.82 points more than the grades of those who did not receive a tutor.

For those familiar with regression analysis: In notation of multivariate regressions, the difference-in-differences estimator is represented by the interaction term between the treatment group and the post-treatment period.

$$Y_{it} = \alpha + \beta_1 T_i + \beta_2 post_t + \beta_3 T_i * post_t + \epsilon_{it},$$

where Y_{it} represents the variable of interest for individual i in period t, T_i is a binary variable indicating whether or not individual i participated in the program, and $post_t$ is a binary variable indicating the period following the program. β_3 is the difference-in-differences estimator and ϵ_{it} represents the error term.

In essence, the difference-in-differences estimation uses the change over time for the untreated group as the counterfactual for the change over time of the treated group. That is, it controls for all the characteristics that do not change over time (both observable and unobservable) and for all the changes over time that affect the treated and untreated group in the same manner.

The key assumption is that without the program, the change over time would have been the same in both groups. This is often referred to as the common or ***parallel trend assumption***. If in the absence of the program, the treated group would have had a different trend over time than the comparison group, this assumption is violated (see Meyer (1995) for a discussion of the parallel trend assumption). These concepts are graphically illustrated in Figure 3.

[Figure 3]

In the case of the student tutoring example, the assumption implies that without the additional help, the children with a tutor and those without one would have improved their scholarly achievements at the same rate. It is not obvious that this is the case here. Even without the program, the children who were originally behind – and were, therefore, more likely to receive a tutor – might have improved more than the other children, given that they had more room to improve. On the other hand, since these children had a

harder time learning, it is also possible that they would have fallen even further behind. The difference-in-differences estimate could in this case be biased upward or downward. This is not possible to assess from the data since we do not know how much the children with a tutor would have improved without a tutor. That is, we cannot test the parallel trend assumption.

In recent studies, researchers have increasingly tried to look at longer time series to see whether the treatment and control groups evolved in parallel before the start of the treatment. This is illustrated in Figure 4. It shows a case in which the treatment group and control group have a parallel trend prior to the treatment. After the treatment starts, the two groups diverge. The finding of a parallel trend before and a difference after treatment gives credibility to the conclusion that the treatment caused the effect.

[Figure 4]

5.1 *Experiences of difference-in-differences in public economics*

Duflo (2001) provides a great illustration of the application of difference-in-differences estimation in practice. The paper takes advantage of variation in school construction in Indonesia across regions and time to measure the impact of school construction on school attendance. It illustrates well how many assumptions need to be taken into account when conducting this type of estimation in a reliable manner.

On the topic of tax administration, Naritomi (2015) uses a difference-in-differences approach to study the effectiveness of incentives for final consumers to ask firms for a receipt. She compares declared revenues of retail versus wholesale firms, before and after the policy change. Providing consumers with a financial incentive to ask for a receipt proves to be effective in boosting firms' declared sales and taxes. Incentives in the form of lotteries seem to be particularly effective, suggesting that consumers might be affected by behavioral biases. These behavioral biases make incentives in the form of lotteries more cost-effective for the government. However, they also raise ethical questions as to whether it is legitimate for the government to exploit such biases.

Casaburi and Troiano (2015) study the electoral response to a nationwide anti-tax evasion policy in Italy using a difference-in-differences estimator. By comparing municipalities with more or less "intensity" of the anti-tax evasion intervention before and after the program, they find that a higher intensity of the program lead to significantly higher re-election chances for the local mayor. There is also a large literature in taxation, particularly focusing on the US and other highly developed countries, using difference-in-differences estimation to analyze the impacts of tax changes on individual behavior, such as labor supply and on firm behavior, such as investment. Reviewing this literature is beyond the scope of this paper.

Bandiera et al. (2009) apply a version of difference-in-differences estimation to study the behavior of public officials in public procurement processes. They exploit a natural experiment in Italy's public procurement system to look at the determinants of waste and inefficiencies in the procurement process. Public entities in Italy can procure goods either directly from providers or from a central platform, where goods are available at pre-negotiated conditions. The study exploits the fact that certain goods were only available on the central platform at certain times. By comparing the price of procured goods during times when goods were available on the central platform to times when they were not, the authors can disentangle the mechanism through which waste happens. The availability of products on the central platform has significant effects on procurement behavior and prices. The results show that this variation in prices is mostly due to passive behavior by the public agents rather than active benefit-seeking, and the effect varies with different governance structures.

A recent study by Lewis-Faupel et al. (2014) also uses difference-in-differences estimation to study public procurement. The study exploits regional and time variation in the adoption of electronic procurement systems across India and Indonesia to test the effect of e-procurement on the cost and quality of infrastructure provision. The fact that both countries rolled out the treatment gradually by region allowed the authors to carry out a difference-in-differences strategy, comparing states that were treated first with those that followed later. They find no effect on the prices paid by the government, but significant improvement in quality.

5.2 *Summary difference-in-differences analysis*

Difference-in-differences analysis compares the change in outcomes over time of those that participated in the program to the change over time of those that did not. The change for those who do not participate in the program represents the counterfactual of the change for those that did participate in the program. The key assumption of this method is the assumption of common trends. It assumes that without the program, both groups would have had identical trajectories over time. The benefit of this method is that it controls for all the characteristics that do not change over time (both observable and unobservable) and for all the changes over time that affect the treated and untreated group in the same manner. The drawback is that it is typically impossible to assess whether the two groups would have developed in the same way in the absence of the program. If this is not the case, the analysis will be biased. When longer time series of date are available, the assumption can be tested to some degree by showing that over a long pre-treatment period, the two groups had the same changes over time, and only when the treatment started did the time trends of the two groups diverge.

6. Matching Procedures and Propensity Scores

Matching procedures are based on the original objective of constructing a representation of the counterfactual and attempting to create a control group that is as similar as possible to the treatment group. There are several matching methods. In the basic case, each individual in the treated group is matched to an individual from the untreated group with the same observable characteristics. The comparison group is then composed of these matched individuals. To estimate the impact of a program, the method compares the outcomes between the treatment group and the matched comparison group. Given that both groups have the same observable characteristics before the program, it is expected that any difference after the program will be due to having been exposed to the program.

We can look at this process in the case of the tutoring program example. It is possible to find children who did not sign up for the program but had the same grades on average as children who received

the help of a tutor before the intervention. This way, a comparison group can be created with non-treated students that have the same observable characteristics as the treated children.

[Figure 5]

Figure 5 shows the matching process for the tutoring example with three characteristics: age, pre-test score, and gender. Students in the treatment group are matched to children who did not receive a tutor. The matched students from the non-treated list then serve as the comparison group. The process of finding similar peers ensures that the two groups are identical along the observable characteristics that are considered for the match.

The key assumption, in this case, is that those who do not participate are, on average, identical to their matched peers, except for having participated in the program. The challenge is that matching can never control for *unobserved* variables. In the tutoring program example, there is a non-random reason that two children with the same grades received a different treatment. Maybe the teacher knew that some students had more potential than others, or maybe some students had more proactive parents who were pushing for their child to receive a tutor. If there are such differences that the available data cannot measure, the selection bias problem arises again, even though on observed characteristics, the two matched groups are identical. It is likely, for example, that in the absence of the tutoring program, children with more proactive parents would have improved more than their classmates with the same grades.

In this context, the benefits of randomized treatment assignment become apparent. Randomized assignment ensures that the treatment and comparison groups are similar not only along observable but also along unobserved characteristics.

The larger the number of characteristics that are included in the matching, the harder it is to use one-to-one matching. With many observed characteristics, it may be impossible to find an identical student that did not have a tutor. For these reasons ***"Propensity Score Matching" (PSM)*** was developed. PSM allows matching with many characteristics. Based on the observable characteristics of individuals, their

propensity (or probability) of being in the treated group is estimated. In this way, the number of characteristics is reduced to a single score, ranging from zero to one, which predicts the probability of participating in the program. In effect, the propensity score is a weighted average of the included characteristics. The matching is then done between individuals that have the same score: that is, the same likelihood of participating in the program. For a detailed guide for implementing matching techniques see Imbens (2015).

6.1 *Experiences of matching in public economics*

One of the earliest and most well-known examples of the propensity score matching technique was conducted by Dehejia and Wahba (1999) to analyze the impact of a labor training program on income. Comparing the propensity score method to other approaches, they find that the propensity score estimate was, in this case, much closer to the results of the randomized experiment than many of the other non-experimental estimates. In this context, Angrist and Pischke (2009) argue that what matters most is including the right covariates, not the type of matching methodology. In the setting studied by Dehejia and Wahba, including the 2-year lagged pre-treatment income turned out to be decisive.

6.2 *Summary of matching*

Matching methods compare outcomes of treated individuals with those of similar individuals that were not treated. In exact matching, participants are matched with individuals that are identical along selected characteristics but that did not participate in the treatment. In propensity score matching, participants are compared to individuals that had the same probability of participating in the program according to their observable characteristics but did not participate. The key assumption of this method is that those who participate in the program are, on average, identical to their matched peers, except for having participated in the program. It assumes that when people or entities are matched on observable characteristics, they will also be comparable along unobserved dimensions. The benefit of this method is that it controls for observed characteristics. The drawback is that it is typically impossible to rule out that there are not also other, unobserved characteristics that differ between the groups, which would bias the impact estimation. Knowing

the likelihood that unobservable characteristics will be important in a given context requires understanding the mechanism by which the participants were selected into the program and knowing what factors other than the program may affect the outcome.

7. Regression Discontinuity Design

Regression discontinuity design (RDD) is a methodology that allows making causal conclusions that are nearly as reliable as the randomized control trial. It can only be applied in cases where a program or policy has a specific threshold that determines who is eligible to participate. A RDD uses the fact that the individuals or entities just barely above the threshold are basically identical to individuals just below. Under certain assumptions, it is therefore possible to measure the treatment effect in the difference between the outcomes of the individuals just below the threshold – who are therefore not eligible – and the outcomes of those just above – who are therefore eligible.

A good illustrative example is a case in which test scores determine whether a student gets admitted to a prestigious college. Imagine that the threshold for being admitted is 924 out of 1,000 possible points. Students who scored 923 points are almost indistinguishable from students with 924 points, but the latter are admitted and the former are not. If the students with 924 or 925 points end up earning much more than the students with 922 or 923 points, this difference may be the result of attending the prestigious college.

For an example in tax administration, assume that a tax authority sends a notification letter to all firms whose declared tax filings indicate a large discrepancy between their self-reported revenue and information about their revenue from third party sources. The tax authority, therefore, suspects these firms of cheating. However, the tax authority does not want to send out too many notifications and decides to send notifications to all firms with discrepancies that are greater than 1,000 dollars. That is, whether a firm receives a notification is determined by whether it has more or less than 1,000 dollars in discrepancies. The regression discontinuity design will then compare firms that had discrepancies a bit smaller than 1,000 dollars to firms that had discrepancies just a bit larger than this cut-off.

Figure 6 displays this example of a regression discontinuity evaluation. The solid line represents the relationship between the size of the discrepancy and the declared tax amount: firms with larger discrepancies also tend to declare more tax. This is likely due to the fact that they are larger in size. Taxpayers above the cutoff value (in our example 1,000 dollars in discrepancies) are included in the treatment, i.e., they receive a notification. Under certain assumptions, the sharp increase in the amount of declared taxes above the cutoff can then be attributed to the notification.

[Figure 6]

The key assumption in a regression discontinuity design is that the entities or individuals just below the cutoff are not systematically different from those just above. This assumption can be violated, for example, if there is strategic manipulation around the threshold. If, for instance, it is known prior to the mailing of the notifications that they will be sent to all firms with more than 1,000 dollars in discrepancy, then firms might be able to manipulate their discrepancy to be just below that cut-off. Those who do so may be particularly shrewd, well-informed, or otherwise different than those who do not. In that case, there will be a difference between the firms just below the threshold and those just above.

Such a difference around the threshold introduces selection bias. The good news is that the assumption that there is no such behavior around the threshold can be tested. If a manipulation occurred, there would be a higher concentration of firms (bunching) just below the threshold. This can be verified. In the same manner, it is possible to verify that there are no differences in the key characteristics between the firms just above or below the threshold.

Finally, a regression discontinuity design also requires that no other programs or policies be applied to the same threshold. For example, if the firms with discrepancies greater than 1,000 dollars are also visited by an auditor, it would not be possible to distinguish the impact of that visit from the impact of the notification. Knowing whether other things change at the same threshold requires good knowledge of the institutional details and the context in which the intervention takes place.

Both problems, the behavioral response to the threshold and the possibility that other policies are applied to the same threshold, are more frequent when the cutoff is known by everyone. Therefore, optimal thresholds for the use of this methodology are often secret or defined only after the score for each individual or entity has already been determined.

One limit of RDDs is that the estimation can only be applied to observations around the cutoff. It is not possible to know what the impact was for firms with discrepancies much larger than 1,000 dollars, or what it would have been for firms with much smaller discrepancies. How informative the insights of the RDD are will therefore depend on the context of the policy and on the extent to which we think that the program affects people or entities that are far away from the threshold differently.

7.1 *Experiences of regression discontinuity in public economics*

RDDs are of particular interest for impact evaluations in the domain of public economics since many policies related to public economics are organized around cut-offs. In tax administration, for instance, there are many policies that are applied according to some cutoff, and frequently the administrative data required for the analysis already exists. Similarly, audit rules for public procurement, tax evasion, labor laws, etc., often use scoring rules with a cut-off, above which entities have a higher risk of being audited.

In an ongoing study, we apply this method to procurement practices in Chile (Gerardino, Litschig and Pomeranz 2015). In collaboration with the national comptroller agency "Contraloría," we exploit a scoring rule that creates higher audit probabilities for public entities above certain thresholds. The study then analyzes the impacts of audits on the public procurement process by comparing public entities that fell just below the cutoff to entities that were just above.

7.2 *Summary of regression discontinuity designs*

RDDs compare the outcomes of individuals or entities that are just below a threshold that qualifies them for the treatment with the results of those that are just above this threshold (or cut-off). Outcomes of individuals or entities that fall just below the threshold represent the counterfactual of the individuals who

fall just above. The key assumption is that the individuals just above the threshold are otherwise almost identical to those who fall just below. This implies that there is no manipulation around the threshold and that no other policies are applied based on the same cutoff. This is more likely to be the case when the exact threshold is not known ex-ante. RDDs can produce very reliable impact estimations. In public administration, there are many policies that are applied according to some cutoff, and frequently the administrative data required for the analysis already exists. The key weakness of RDDs is that the effect can only be estimated for individuals or entities that are close to the cutoff.

8. Conclusion

Rigorous impact evaluations have experienced a large expansion in recent years, both in their methodological developments and in their practical applications. Public agencies interested in affecting their citizens to encourage behaviors such as tax compliance, savings or rule following, are increasingly testing the effectiveness of public policies to achieve these goals. This paper aims to provide an introductory overview for those interested in conducting such evaluations in a reliable way. Among the methods covered, randomized evaluations and regression discontinuity designs provide the most rigorous, causally valid estimates. If these methods are not available, difference-in-differences estimation or matching methods may provide an alternative. These latter methods are more likely to suffer from selection bias or omitted variable bias and, therefore, have to be applied with more caution. Finally, simple differences and pre-post analysis, while being frequently applied in practice by the media or policymakers due to their conceptual simplicity, are also the most prone to estimation biases and are therefore generally the least reliable of the methods described in this paper.

Apart from the particular method that is chosen, the quality of the evaluation will depend to a large degree on two factors: the quality of the execution of the analysis and detailed knowledge of the context of the program that is being evaluated. This is why the increasing number of collaborations between academics

and practitioners hold so much promise. Combining the methodological knowledge of highly trained academics with the expertise of public officials about the practical context has huge potential to grow our understanding of both public finance and behavioral economics.

9. References

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* 105 (490): 493-505.

Abadie, Alberto. 2005. Semiparametric difference-in-differences estimators. *Review of Economic Studies* 72 (1): 1-19.

Abdul Latif Jameel Poverty Action Lab (J-PAL). 2015. Why Randomize? Case Study. Cambridge, MA: Abdul Latif Jameel Poverty Action Lab. http://www.povertyactionlab.org/methodology/why/why-randomize

Angrist, Joshua D., and Jörn-Steffen Pischke. 2015. *Mastering 'Metrics: The Path from Cause to Effect.* Princeton, USA: Princeton University Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, USA: Princeton University Press.

Angrist, Joshua D., and Victor Lavy. 1999. Using Maimonides' Rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* 114 (2): 533-575.

Ariel, Barak. 2012. Deterrence and moral persuasion effects on corporate tax compliance: Findings from a randomized controlled trial. *Criminology* 50: 27–69.

Bandiera, Oriana, Andrea Prat and Tommaso Valleti. 2009. Active and passive waste in government spending: Evidence from a policy experiment. *American Economic Review* 99 (4): 1278-1308.

Banerjee, Abhijit V., and Esther Duflo. 2009. The experimental approach to development economics. *Annual Reviews of Economics* 1: 151-178.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo and Leigh Linden. 2007. Remedying education: Evidence from two randomized experiments in India. *Quarterly Journal of Economics* 122(3): 1235-1264.

Bhargava, Saurabh and Dayanand Manoli. 2015. Psychological Frictions and the incomplete take-up of social benefits: Evidence from an IRS field experiment. *American Economic Review* 105 (11): 1-42.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics* 119 (1): 249-275.

Blumenthal, Marsha, Charles Christian and Joel Slemrod. 2001. Do normative appeals affect tax compliance? Evidence from a controlled experiment in Minnesota. *National Tax Journal* 54: 125–36.

Carrillo, Paul, Dina Pomeranz and Monica Singhal. 2014. Dodging the taxman: Firm misreporting and limits to tax enforcement. NBER Working Paper #20624.

Carroll, Gabriel D., James J. Choi, David Laibson, Brigitte Madrian and Andrew Metrick. 2009. Optimal defaults and active decisions. *Quarterly Journal of Economics* 124 (4): 1639-1674.

Casaburi, Lorenzo and Ugo Troiano. 2015. Ghost-house busters: The electoral response to a large anti-tax evasion program. NBER Working Paper #21185.

Castro, Lucio, and Carlos Scartascini. 2015. Tax compliance and enforcement in the Pampas. Evidence from a field experiment. *Journal of Economic Behavior and Organization* 116: 65-82.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2014. Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review* 104 (9): 2633–2679.

Chetty, Raj. 2015. Behavioral economics and public policy: A pragmatic perspective. *American Economic Review* 105(5): 1-33.

Choi, James J., David Laibson, Brigitte C. Madrian, and Andrew Metrick. 2003. Optimal Defaults. *American Economic Review* 93: 180–185.

Coleman, Stephen. 1996. The Minnesota income tax compliance experiment: State tax results. Munich Personal RePec Archive Paper No. 4827, University of Munich.

Dehejia, Rajeev H., and Sadek Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94 (448): 1053-1062.

Del Carpio, Lucia. 2013. Are the neighbors cheating? Evidence from a social norm experiment on property taxes in Peru. Princeton University Working Paper, Princeton, NJ.

Duflo, Esther, Rachel Glennerster and Michael Kremer. 2008. Using randomization in development economics research: A toolkit. *Handbook of Development Economics* 4: 3895-3962.

Duflo, Esther. 2001. Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment. *American Economic Review* 91: 795-813.

Dupas, Pascaline and Jonathan Robinson. 2013. Why don't the poor save more? Evidence from health savings experiments. *American Economic Review*, 103(4): 1138–1171.

Dwenger, Nadja, Henrik J. Kleven, Imran Rasul and Johannes Rincke. 2014. Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in Germany. Working Paper, Max Planck Institute for Tax Law and Public Finance, Munich.

Fellner, Gerlinde, Rupert Sausgruber and Christian Traxler. 2013. Testing enforcement strategies in the field: threat, moral appeal and social information. *Journal of the European Economic Association* 11 (3): 634–60.

Gangl, Katharina, Benno Torgler, Erich Kirchler and Eva Hoffmann. 2014. Effects of supervision on tax compliance. *Economics Letters* 123 (3): 378–82.

Gerardino, Maria Paula, Stephan Litschig, and Dina Pomeranz. 2015. Monitoring public procurement: Evidence from a regression discontinuity design in Chile. Mimeo.

Gerber, Alan S., and Donald P. Green. 2012. *Field experiments: Design, analysis, and interpretation*. New York, USA: WW Norton.

Gertler, Paul, Sebastian Martinez, Patrick Premand, Laura B. Rawlings and Christel M. J. Vermeersch. The World Bank. 2011. *Impact Evaluation in Practice*. Washington, D.C: World Bank Group. http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1295455628620/Impact_Evaluation_in_Practice.pdf.

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton, USA: Princeton University Press.

Glewwe, Paul, Michael Kremer and Sylvie Moulin. 2009. Many children left behind? Textbooks and test scores in Kenya. *American Economic Journal: Applied Economics* 1(1): 112-35.

Hallsworth, Michael, John A. List, Robert D. Metcalfe and Ivo Vlaev. 2014. The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. NBER Working Paper #20007.

Hallsworth, Michael. 2014. The use of field experiments to increase tax compliance. *Oxford Review of Economic Policy* 30 (4): 658–679.

Harju, Jarkko, Tuomas Kosonen, and Oli Ropponen. 2013. Do honest hairdressers get a haircut? On tax rate and tax evasion. Government Institute for Economic Research Finland Working Paper, Helsinki.

Hasseldine, John, Peggy Hite, Simon James and Marika Toumi. 2007. Persuasive communications: Tax compliance enforcement strategies for sole proprietors. *Contemporary Accounting Research* 24(1): 171–94.

Haynes, Laura C., Donald P. Green, Rory Gallagher, Peter John, and David J. Torgerson. 2013. Collection of delinquent fines: An adaptive randomized trial to assess the effectiveness of alternative text messages. *Journal of Policy Analysis and Management* 32(4): 718-730.

Imbens, Guido W., and Thomas Lemieux. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142 (2): 615-635.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47 (1): 5-86.

Imbens, Guido W. 2015. Matching papers in practice: Three examples. *Journal of Human Resources* 50 (2): 373-419.

Iyer, Govind S., Philip M. J. Reckers and Debra L. Sanders. 2010. Increasing tax compliance in Washington State: A field experiment. *National Tax Journal* 63 (1): 7–32.

Karlan, Dean and Jonathan Zinman. 2014. Price and control elasticities of demand for savings. Yale University Working Paper.

Karlan, Dean, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman. 2010. Getting to the top of mind: How reminders increase saving. NBER Working Paper #16205.
Kast, Felipe, and Dina Pomeranz. 2014. Saving more to borrow less: Experimental evidence from access to formal savings accounts in Chile. NBER Working Paper #20239.
Kast, Felipe, Stephan Meier, and Dina Pomeranz. 2014. Under-Savers anonymous: Evidence on self-help groups and peer pressure as a savings commitment device. NBER Working Paper #18417.
Kleven, Henrik J., Martin B. Knudsen, Claus T. Kreiner, Soren Pedersen and Emmanuel Saez. 2011. Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica* 79 (3): 651–692.
Lalonde, Robert J. 1986. Evaluating the econometric evaluations of training programs. *American Economic Review* 76: 604-620.
Lee, David S., and Thomas Lemieux. 2010. Regression discontinuity designs in economics. *Journal of Economic Literature* 48 (2): 281–355.
Lewis-Faupel, Sean, Yusuf Neggers, Benjamin A. Olken and Rohini Pande. 2014. Can electronic procurement improve infrastructure provision? Evidence from public works in India and Indonesia. NBER Working Paper #20344.
Litschig, Stephan and Yves Zamboni. 2013. Audit risk and rent extraction: Evidence from a randomized evaluation in Brazil. Barcelona Graduate School of Economics Working Paper, Barcelona.
Ludwig, Jens, Jeffrey Kling, and Sendhil Mullainathan. 2011. Mechanism experiments and policy evaluations. *Journal of Economic Perspectives* 25 (3): 17-38.
Madrian, Brigitte C., and Dennis Shea. 2001. The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior. *Quarterly Journal of Economics*, 116: 1149–118.
Meyer, Bruce D. 1995. Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics* 13 (2): 2. 151-161.
Miguel, Edward and Michael Kremer. 2004. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72 (1): 159-217.
Naritomi, Joanna. 2015. Consumers as tax auditors. London School of Economics Working Paper, London.
Organization for Economic Cooperation and Development (OECD). 2010. Understanding and influencing taxpayers' compliance behaviour. OECD, Paris.
Ortega, Daniel and Pablo Sanguinetti. 2013. Deterrence and reciprocity effects on tax compliance: Experimental evidence from Venezuela. Development Bank of Latin America, Caracas.
Pomeranz, Dina, Cristobal Marshall, and Pamela Castellon. 2014. Randomized tax enforcement messages: A policy tool for improving audit strategies. *Tax Administration Review*, 36 (1): 1–21.
Pomeranz, Dina. 2015. No taxation without information: Deterrence and self-enforcement in the value added tax. *American Economic Review* 105 (8): 2539-69.
Prina, Silvia. 2015. Banking the poor via savings accounts: Evidence from a field experiment. *Journal of Development Economics* 115: 16-31.
Ries, Eric. 2011. *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.* New York, USA: Crown Business Inc.
Slemrod, Joel, Marsha Blumenthal and Charles Christian. 2001. Taxpayer response to an increased probability of audit: Evidence from a controlled experiment in Minnesota. *Journal of Public Economics* 79 (3): 455–83.
Torgler, Benno. 2013. A field experiment on moral suasion and tax compliance focusing on under-declaration and over-deduction. *Public Finance Analysis* 69 (4): 393-411.
Torgler, Benno. 2004. Moral suasion: An alternative tax policy strategy? Evidence from a controlled field experiment in Switzerland. *Economics of Governance* 5 (3): 235–53.
Wenzel, Michael and Natalie Taylor. 2004. An experimental evaluation of tax-reporting schedules: A case of evidence-based tax administration. *Journal of Public Economics*, 88 (12): 2785–99.
Wenzel, Michael. 2006. A letter from the tax office: Compliance effects of informational and interpersonal justice. *Social Justice Research* 19 (3): 345–64.

Wenzel, Michael. 2005. Misperceptions of social norms about tax compliance: From theory to intervention. *Journal of Economic Psychology* 26 (6): 862–83

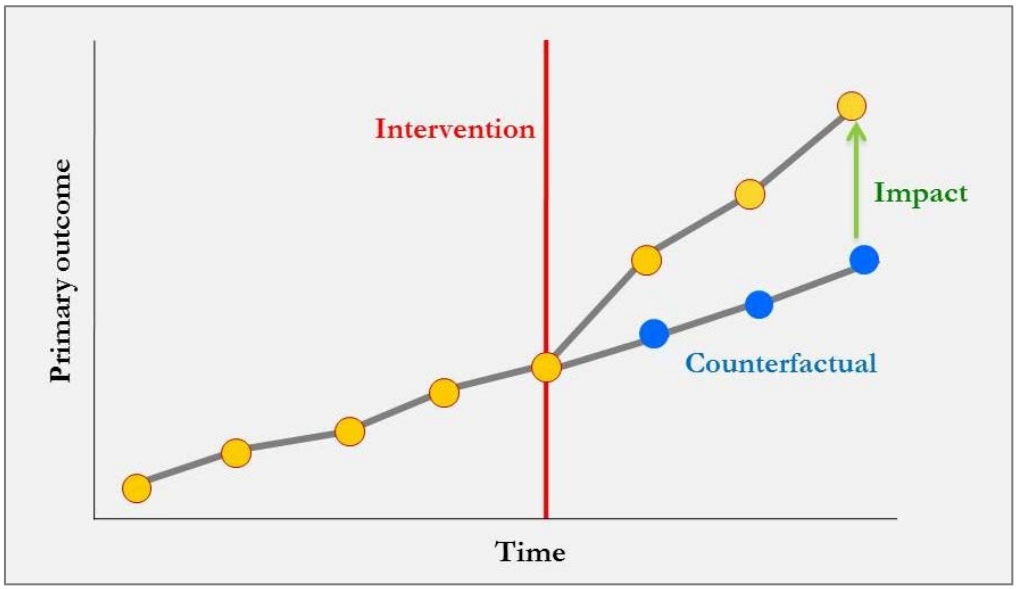
Tables and Figures

Table 1. Estimating Difference-in-Differences

	Result before the program	Result after the program	Difference over time
Treated group	24.80	51.22	26.42
Untreated group	36.67	56.27	19.60
Difference-in-differences estimate			6.82

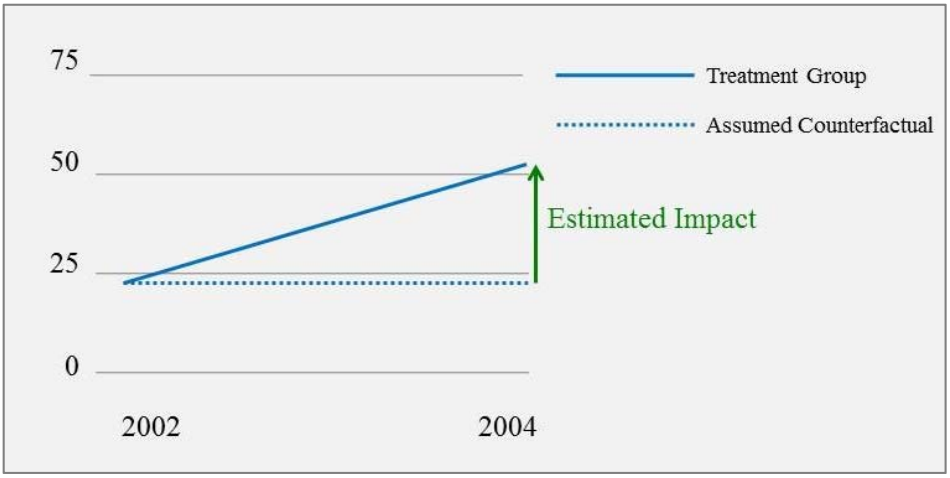
This table provides a numerical example a difference-in-differences estimation. The numbers are from the tutoring example and represent grades of the children with and without the tutoring program, before and after the program. Source: Abdul Latif Jameel Poverty Action Lab (2015).

Figure 1. Counterfactual



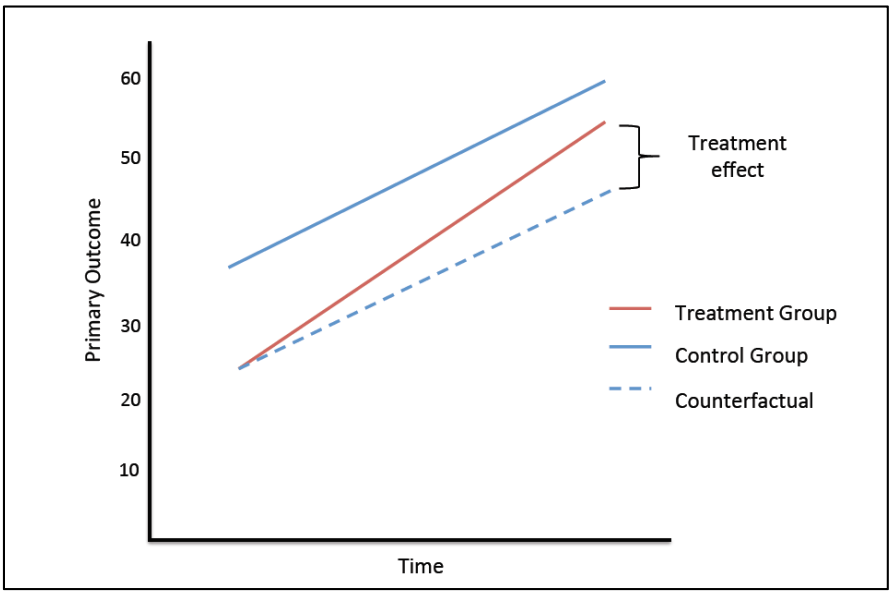
This figure represents the fundamental challenge of impact evaluations, which seek to measure the difference between the outcome that occurred (shown in light/yellow dots) and a counterfactual that is never observed (shown with dark/blue dots). Impact evaluation techniques therefore – implicitly or explicitly – attempt to construct an estimation of the counterfactual in order to measure the impact. This is often done through the use of a control group. Source: Abdul Latif Jameel Poverty Action Lab (2015).

Figure 2. Counterfactual Assumption for Pre-Post: No Change in the Absence of Treatment



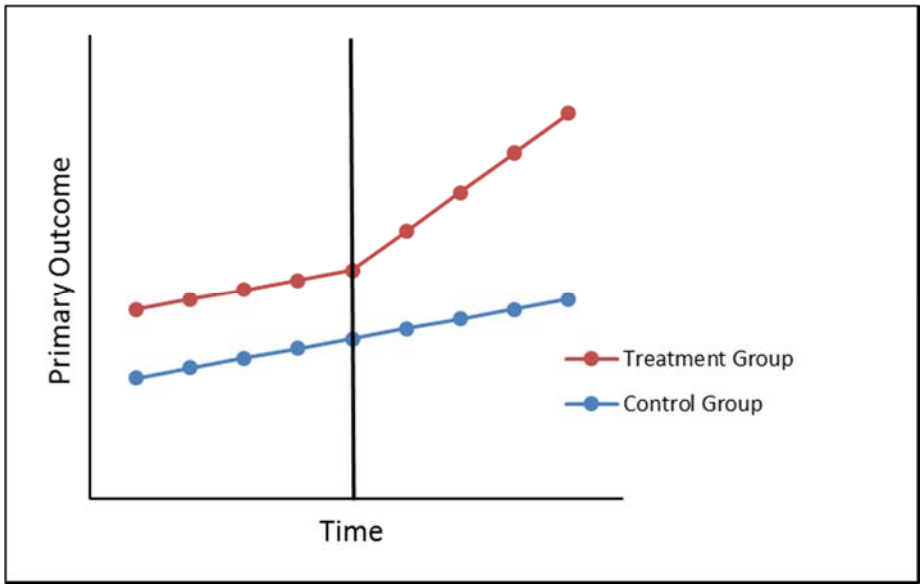
In a pre-post impact evaluation, the key assumption is that in the absence of the treatment, there would have been no change in the outcome variable. If this is the case, the pre-treatment situation represents a valid counterfactual for the post-treatment situation. Source: Abdul Latif Jameel Poverty Action Lab (2015).

Figure 3. Counterfactual Assumption in Difference-in-Differences Analysis: Parallel Trends



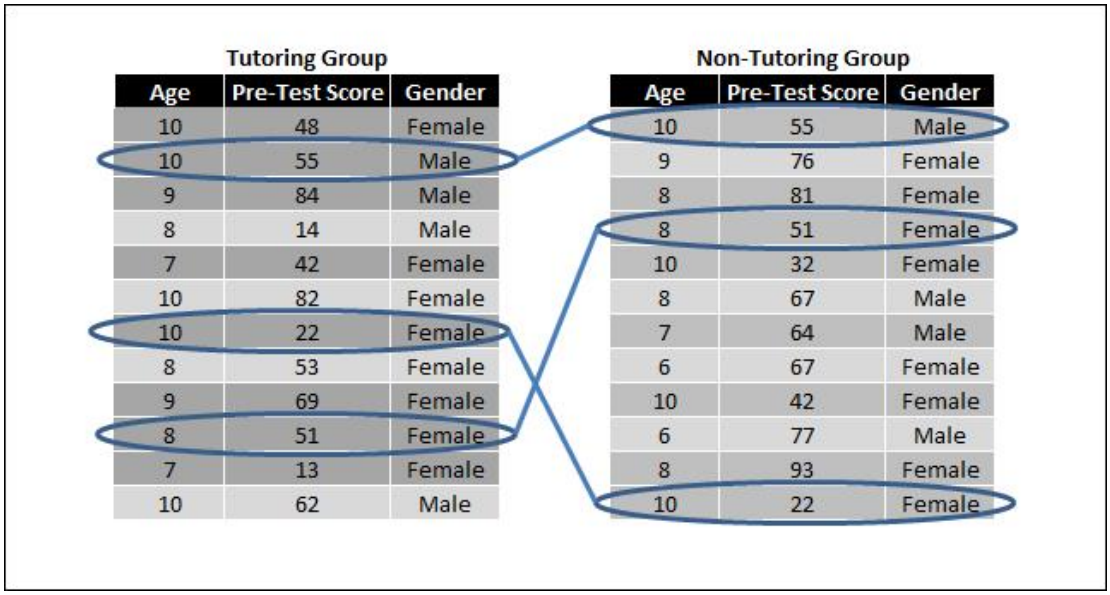
This figure displays the logic and assumptions underlying the difference-in-differences analysis. The counterfactual of the change over time for those that did participate in the program is the change for those that did not participate (represented by the dashed line). The key assumption is therefore that in the absence of the treatment, the two groups would have followed the same trend over time. If this is true, the treatment effect can be measured as the difference between the differences over time. See also Table 1. Source: Abdul Latif Jameel Poverty Action Lab (2015).

Figure 4. Checking for Parallel Trends



This figure demonstrates how time series data can allow us to check for parallel trends between the treatment group and the control group prior to the start of the treatment. As the figure shows, prior to the treatment, the two groups have a parallel trend. This gives credibility to the assumption that in the absence of treatment, they would have continued on a parallel trend, which is required for the difference-in-differences analysis to be valid. The two groups diverge only after treatment, giving credibility to the conclusion that the treatment led to this effect.

Figure 5. Matching Process in the Tutor Example



Tutoring Group

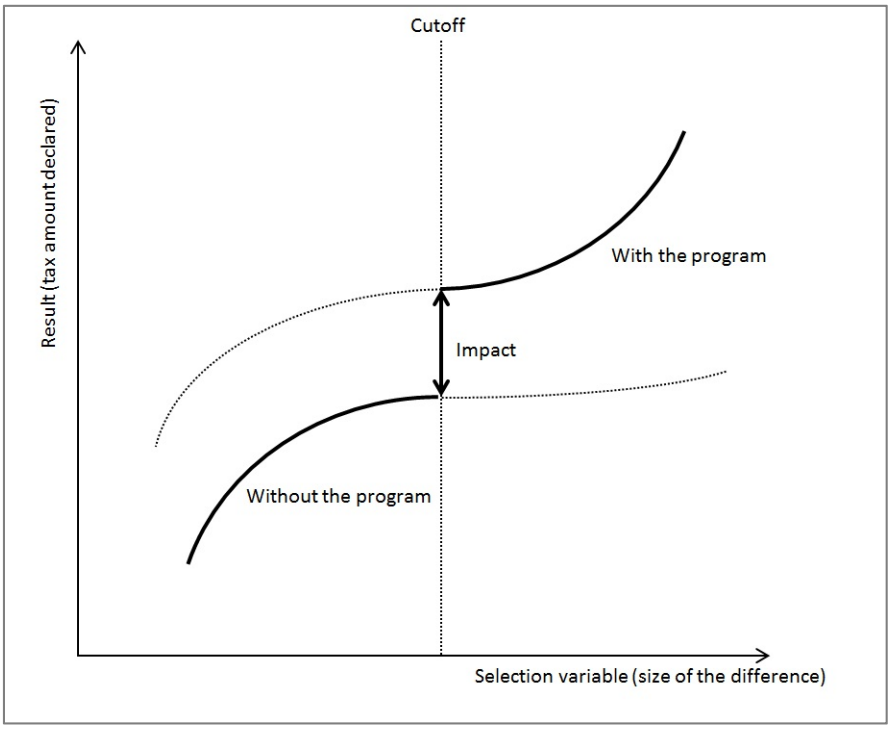
Age	Pre-Test Score	Gender
10	48	Female
10	55	Male
9	84	Male
8	14	Male
7	42	Female
10	82	Female
10	22	Female
8	53	Female
9	69	Female
8	51	Female
7	13	Female
10	62	Male

Non-Tutoring Group

Age	Pre-Test Score	Gender
10	55	Male
9	76	Female
8	81	Female
8	51	Female
10	32	Female
8	67	Male
7	64	Male
6	67	Female
10	42	Female
6	77	Male
8	93	Female
10	22	Female

This is an example of a direct matching process for the tutoring example. It matches students in the treatment group to students who did not receive a tutor. Matching is done along three observable dimensions: age, pre-test-score, and gender. The matched students from the non-treated list then serve as the comparison group. Source: Abdul Latif Jameel Poverty Action Lab (2015).

Figure 6. Illustration of a Regression Discontinuity Design



This figure provides a graphical representation of a RDD. Individuals or entities above a certain cutoff of the selection variable are included in the treatment, and those below the cutoff are not. That is, there is a discontinuity along the selection variable, above which the treatment is applied. If the required assumptions for a RDD are met, the sharp increase in the outcome variable at the cutoff can be attributed to the treatment. Source: Abdul Latif Jameel Poverty Action Lab (2015).